跨学科视角下基因工程领域热点交叉主题识别及主题演化分析

朱世琴[1] 范丹丹 郭田雨

华东理工大学科技信息研究所，上海，200237

摘 要：为了更加精准把握交叉学科研究热点与发展态势，本研究提出一种计算主题学科交叉度的方法，并结合主题强度来综合识别热点交叉主题、对交叉主题未来发展进行预测。本研究选取 Web of Science 数据库 2000-2019 年基因工程领域论文进行实证分析，首先采用 LDA 模型挖掘主题，然后通过计算主题强度和主题学科交叉度识别热点交叉主题，最后划分时间窗口，绘制主题强度及主题学科交叉度的变化趋势图并对结果展开分析。实证结果表明：基因工程领域共有 21 个重要主题，其中 7 个热点主题，14 个学科交叉主题，2 个热点交叉主题；根据主题强度变化趋势，将 21 个主题划分为 3 个上升型主题，7 个下降型主题和 11 个平稳型主题，大部分主题的学科交叉程度呈现上升趋势。

关键词：学科交叉主题；热点主题；主题识别；主题演化

中图法分类号：G353.1

Theme Evolution Analysis And Recognition Of Hot Interdisciplinary Themes In Genetic Engineering From An Interdisciplinary Perspective

Zhu Shiqin Fan Dandan Guo Tianyu

East China University of Science and Technology, Shanghai, 200237

Abstract: In order to more accurately grasp the hot spots and development trends of interdisciplinary research, this study proposes an integrated approach based on theme intensity and theme interdisciplinary degree to identify hot interdisciplinary themes and predict the future development of interdisciplinary themes. In this study, papers in the field of genetic engineering from 2000-2019 in Web of Science database were selected for empirical analysis. Firstly, themes were mined using LDA model, then hot interdisciplinary themes were identified by calculating theme intensity and interdisciplinary degree, and finally, time windows were divided to plot variation trend of theme intensity and interdisciplinary degree and the results were analyzed. The empirical results show that there are 21 important themes in the field of genetic engineering, including 7 hot themes, 14 interdisciplinary themes, and 2 hot interdisciplinary themes. According to the variation trend of theme intensity, the 21 themes are classified into 3 ascending themes, 7 descending themes, and 11 stable themes, and the interdisciplinary degree of most themes shows an increasing trend.

---

[1]作者简介：朱世琴，硕士，研究馆员，硕士生导师，研究方向为信息计量、分析与评价，Email：shqzhu@ecust.edu.cn；范丹丹，硕士研究生；郭田雨，硕士研究生。

Keywords: Interdisciplinary theme; Hot theme; Theme recognition; Theme evolution

引言

1986 年，诺贝尔基金会主席在颁奖致辞中表示：在物理学和化学之间、生物学和医学之间，旧的学科界限已在各个方面被突破，它们不仅互相交叉，而且形成了没有鲜明界限的连续区。近年来，一系列科学发现、科技创新成果广泛分布在分子生物学、物理化学、系统科学等交叉领域。未来可以预见，在国家重大战略需求的驱动下，多学科交叉会聚与多技术跨界融合将成为常态，并不断催生新学科前沿、新科技领域和新创新形态[1]。

学科交叉主题是两个或多个学科在相互融合、渗透的过程中，形成的共同研究主题，是知识融合的汇聚点[2]、知识扩散的枢纽点[3]、也是实现科技创新的突破点[4]。热点主题是科研工作者在某段时间保持高度关注并展开大量研究的重要主题领域。利用科学方法识别研究热点及演化趋势，能够帮助科研工作者正确把握现实动态，透视学科发展和学术进步，厘清学科重点[5]。目前，学科交叉研究主要分为宏观和微观两个方面[6]，本文将从微观的角度，以基因工程领域为例，通过 LDA 模型提取主题，提出主题学科交叉度的测度方法，并结合主题强度指标识别热点交叉主题；定量分析领域内主题强度及交叉度变化趋势特征，有助于把握交叉学科发展态势，发掘创新性的研究方向和主题。

1 相关研究回顾

1.1 学科交叉主题识别研究

学科交叉主题的识别通常采用引文分析法、词汇分析法和主题模型法进行。

（1）引文分析法。引文分析法通常基于引用关系，对期刊、文献、主题、作者等各类分析对象间的引证现象进行分析。在学科交叉主题识别领域，也可以用引文分析法来识别交叉主题。Chi R 等[7]基于共被引网络分析，发现主要研究主题的发展以及它们之间的关系。Adams J 等[8]在艾滋病研究领域通过构建书目耦合网络，来识别主题内容、确定主题聚类情况。P. Vugteveen 等[9]根据期刊引文关系，绘制交叉学科河流学的学科地图和知识流，并根据引用文献的相似性对论文进行聚类获得研究主题并将其关联到主要学科，但未考虑主题的交叉程度。

（2）词汇分析法。词汇分析法以文献中的词汇作为分析对象，主要采用词频统计和共词分析两种方法来研究热点主题。Xu 等[10]以情报学为例，通过计算 TI 值、Bet 值、词频值等来确定学科交叉主题，将社会网络分析方法与时间序列分析方法相结合来分析学科交叉的演变。学者杜丽君[11]以情报学和计算机科学中与信息检索相关的论文进行研究，通过建立共词矩阵发现两个学科在该领域的交叉研究主题。在基因工程疫苗领域，隗玲[12]等学者采用共词分析方法，利用专利共词聚类和战略坐标图识别技术主题及其发展现状；罗瑞[13]以主题词共现网络表征知识网络，并用结构熵对知识网络状态进行测度，以便进一步识别科学突破主题。

（3）主题模型法。应用到学科交叉主题识别中的主题模型主要有 CTM 模型、AT 模型、LDA 模型等。主题模型的核心计算问题就是利用可视的文档来推断其隐含的主题结构[14]。

潜在狄利克雷分配模型（Latent Dirichlet Allocation, LDA）是由 D.Blei[15]在 2003 年提出的一种常见的主题模型，在主题识别领域应用广泛，例如：张斌[16]运用 LDA 模型从聚类角度探究了混合学科研究主题的形成。陈琼等[17]利用 LDA 模型识别和划分医学信息学领域主题，随后引入 DIV 测度指标比较学科交叉态势。韩正琪等[18]使用 Rao-Striling 指标发现学科交叉程度较高的文献，运用 LDA 模型获取纳米科技领域的高学科交叉文献的研究主题，但未考虑主题的学科交叉度。

1.2 学科交叉态势演化研究

学科交叉研究在国内外学术界引起了广泛的关注与讨论，学科交叉态势演化研究正处于蓬勃发展阶段，其研究对象主要是期刊和学科领域。

（1）期刊为研究对象。Silva[19]、Leydesdorff[20]等学者通过构建引文网络来衡量学科间的差异性。孟祥保[21]对国外图书情报学核心期刊进行研究，了解国外图书情报学学科交叉融合现状，发现其知识来源与应用。R.Agarwal[22]、杨瑞仙[23]均以期刊为研究对象, 从参考和引证两个角度切入研究, 前者证实了信息系统学科边界在不断扩展，后者研究了图情学科和其他学科间的交叉融合情况。

（2）学科领域为研究对象。Carley 和 Porter[24]使用 Rao 多样性作为学科交叉的度量标准，并分析了六个主题类别的论文集的引文模式。研究发现数学学科交叉性很低，而医学学科交叉程度很高，揭示了学科之间整合的趋势。Levitt 等[25]分析了三个特定年份（1980 年，1990 年和 2000 年）社会科学引文索引（SSCI）类别中学科间的演变。曹嘉君等[26]以人工智能领域为分析对象，揭示领域内核心学科类别分布情况，并通过计算各学科相似性得到学科之间的关联及演变情况，了解人工智能领域内各学科的发展态势。Deng 和 Xia[27]采用社会网络分析法以及学科多样性测度方法，研究发现信息行为领域内的学科分布不均衡。

综上所述，国内外学者们从学科、期刊等角度，对交叉学科主题识别、学科交叉态势演化进行了大量卓有成效的研究。利用主题模型法进行主题识别，可以一定程度上克服引文分析法的滞后性以及传统的共词分析法无法体现词对间语义关联的缺陷，并且主题模型法已运用在医学信息学、纳米科技等学科领域，其适用于交叉学科领域的可行性得以充分证明，但尚未有研究将其应用于基因工程领域。故本文选取能够分析潜在语义信息的 LDA 模型提取研究主题，结合主题强度和本文提出的主题学科交叉度的计算方法，识别热点交叉主题。并且扩展学科交叉态势演化研究的研究对象，以基因工程领域为研究对象，探析该领域交叉主题的变化趋势。

2 研究设计与方法

2.1 研究框架

为识别基因工程领域热点交叉主题并进行主题演化分析，提出研究框架设计如图 1 所示。

首先，获取来源于 Web of Science 基因工程领域的论文集，对数据进行去重、删除缺失值、统计词频、去停用词等操作，用 Python 自然语言提取文献的作者关键词（DE）和扩展关键词（ID）作为主题识别研究的语料来源。其次，利用 LDA 主题模型进行主题挖掘，计

算主题强度和主题学科交叉度并确定阈值，根据二者阈值识别热点交叉主题。最后，从主题强度和主题学科交叉度两个方面进行主题分类和演化趋势呈现，并结合为主题发展做出贡献的学科类别，对主题发展态势做出合理分析。

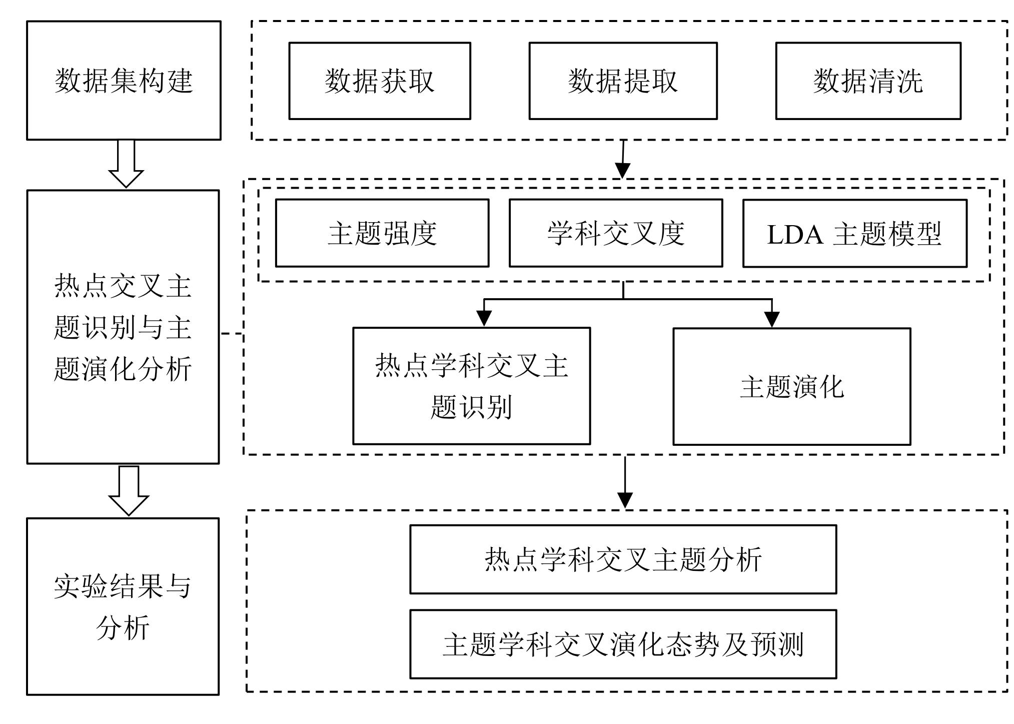


图 1　研究框架设计

2.2 研究方法

2.2.1 基于 LDA 模型的热点交叉主题识别方法

本研究采用 LDA 主题模型用于基因工程领域主题研究具有明显优势：面对数据量庞大的基因工程领域文献，LDA 主题模型方法展现了强大的文本处理能力，能够使用计算机语言实现对文献的作者关键词和扩展关键词进行主题挖掘，提取表现力更强的特征词汇，更加精准地挖掘基因工程领域主题；LDA 主题模型最大的优势是将主题挖掘与主题演化相结合，在获取主题的同时还可以分析主题的演化趋势，把握领域的研究方向。

（1）基于 LDA 模型的主题挖掘

首先对经过预处理的语料创建词语词典，对每个单独的词语赋予一个索引，使用创建的词典，将文档列表转化成矩阵；其次使用 Gensim 模型来建立 LDA 模型对象，并根据计算困惑度获得最优主题数 K，之后在矩阵上运行并训练 LDA 模型，输出主题-词语概率分布矩阵，各词语按照频率依次从大到小输出，选取每个主题下概率排在前 10 的词汇代表该主题，再结合其他输出词汇对主题进行标识，在此过程中生成文档-主题概率分布矩阵及主题-词项概率分布矩阵。

（2）主题强度测度

热点主题挖掘及主题演化可以通过计算主题强度衡量，主题强度可以反映主题的重要程度和关注程度，它通过比较在相同时间窗口下不同主题的主题强度来挖掘热点主题，分析同一个主题在连续、不同的时间窗口下的主题强度变化揭示主题演化~~特征~~趋势。主题强度通过

词汇可能出现的概率进行计算，词汇分布概率是在 LDA 主题模型下通过上下文语境抽取。主题强度计算主要通过构建的文档-主题概率分布矩阵获得每个主题由每篇文档生成的概率。

$$\theta_z^t = \frac{\sum_{d=1}^{D^t} \theta_z^d}{D^t} \tag{1}$$

式中，θ_z^t表示 t 时间段的主题强度，取主题后验概率平均值获得，θ_z^d表示主题 z 占文档 d 的比例，D^t表示在 t 时间段的文档数量。

计算出每个主题的主题强度后，确定一个阈值以便筛选出关注度较高的主题。关于主题强度阈值的确定，本文采用吴查科等[28]提出的主题强度阈值计算方法，计算公式如下：

$$T = \frac{\sum_d \sum_z \theta_z^d}{D^t K} \tag{2}$$

式中，T 为主题强度阈值，K 表示主题的个数，D^t表示文本集合，当主题强度大于阈值 T 时，可以判断该主题为当前时间窗口的热点主题。

（3）主题学科交叉度测度

根据构建的公式对主题的学科交叉度进行测度。主要思路为在得到文档-主题概率分布矩阵后，获得每个主题下包含的文档集，并计算出每篇文档的学科交叉度，本文采用 Rao-Stirling 指标作为学科交叉综合测度指标，在文中简称为 R。R 指标从多样性、均衡性及差异性综合测度单篇论文学科交叉程度。若一篇论文参考文献所属学科类别非常相似，则该论文学科交叉程度较低，反之，则越高。

$$\mathrm{R} = \sum_{i,j} (1 - S_{i,j}) p_x p_j \tag{3}$$

其中 p_x表示学科 i 的被引频次占所有学科总被引频次的比例, $S_{i,j}$表示学科 i 和学科 j 的相似性程度矩阵中学科 x 和学科 j 之间的相似性。

本文提出主题学科交叉度的计算方法，根据一个主题下所有文档的学科交叉度的均值确定该主题的学科交叉度。

$$R_t = \frac{1}{m} \sum_{i=1}^{m} R_i \tag{4}$$

式中，R_t表示第 t 个主题的学科交叉度，m 表示主题包含的文档数，R_i是第 i 篇文档的学科交叉度。

计算出每个主题的主题学科交叉度后，确定一个阈值以便筛选出学科交叉度较高的主题，计算公式如下：

$$\mathrm{I} = \frac{\sum_{t=1}^{K} \mathrm{R_t}}{\mathrm{K}} \tag{5}$$

式中，I 为主题学科交叉度阈值，K 表示主题的个数，R_t表示第 t 个主题的学科交叉度。

2.2.2 主题演化趋势分析方法

主题演化趋势分析包括主题强度变化趋势分析和主题学科交叉度变化趋势分析。现有研究[29]根据引入时间方式的不同，归纳出三种不同演化方法：Joint 法、先离散分析法、后离散分析法。本研究采用后离散分析（Post-discretized Analysis）。这种方法首先忽略了时间，

将整个文本集作为分析文本，通过 LDA 主题模型获得的主题-词项概率分布矩阵以及文档-主题概率分布矩阵，将文档按照其所属时间离散到各时间窗口；最后，通过公式（1）依次计算各主题在连续时间窗口下的主题强度，通过强度上升和下降趋势对主题进行类别划分。

3 实证研究

为了挖掘基因工程领域的热点交叉主题并进行主题演化分析，本文计算获取了各个主题的主题强度和学科交叉度变化趋势，展开了实证研究。

3.1 数据采集与处理

研究数据来源于 Web of Science 的基因工程领域，检索策略如表 1 所示。对数据进行去重、删除无效内容等操作后，最终共得到 51,954 条文献。

表 1　文献检索策略

检索策略	内容
检索式	TI=("gene* engineering" or "DNA engineering" or "gene* manipulat*" or "DNA manipulat*" or "gene* recombinat*" or "transgen*" or "gene* clon*" or "molecular clon*") or AK=("gene* engineering" or "DNA engineering" or "gene* manipulat*" or "DNA manipulat*" or "gene* recombinat*" or "transgen*" or "gene* clon*" or "molecular clon*")
来源数据库	SCI-Expanded 数据库
文献类型	Article
语种	不限
起止时间	2000-2019

其次用 Python 自然语言提取文献的作者关键词（DE）、扩展关键词（ID）、学科、发表时间、标题、摘要等作为待分析文本保存。再次对待分析文本进行词频统计，根据统计结果对不具有区分度的高频词及无意义干扰词进行删除、对同义词进行合并、去停用词、对拥有不同词性的词语进行词形还原。

3.2 热点交叉主题识别

（1）主题抽取

本文使用 Gensim 模型来建立 LDA 模型对象，并根据计算困惑度获得最优主题数 K，2000 年至 2019 年不同主题数目下困惑度分布曲线如下图 2 所示，当困惑度数值波动趋于平缓处于较小值，或出现较为明显的拐点时，则该拐点代表主题模型的拟合程度最好，主题提取效果最佳，因此，确定 K 值为 21。

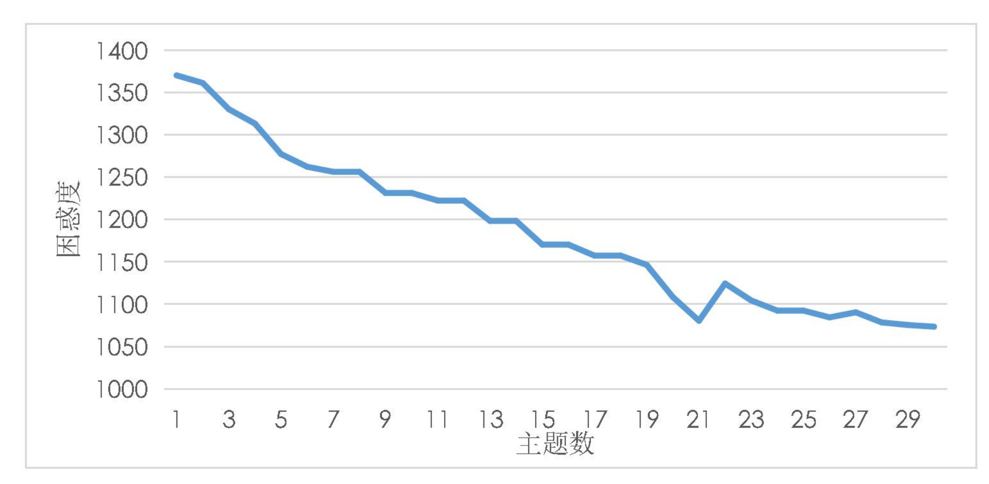


图 2 基因工程不同主题数目下困惑度分布曲线

对语料库运行并训练 LDA 模型，在此过程中生成文档-主题概率分布矩阵及主题-词项概率分布矩阵。输出主题-词项概率分布矩阵，主题下的特征词按对主题的定义程度从大到小排序，在该主题下的分布概率越高，其位置越靠前，选取每个主题下概率排在前 10 的关键词代表该主题，部分主题如图 3 所示，本研究共归纳总结出 21 个主题。

T1 转录组测序技术		T2 转基因药物		T3 转基因作物	
0.066*	"protein"	0.067*	"system"	0.031*	"pcr"
0.062*	"identification"	0.024*	"adaptation"	0.017*	"gene flow"
0.031*	"pathway"	0.018*	"dynamic"	0.015*	"bt corn"
0.023*	"sequence"	0.013*	"biology"	0.011*	"food"
0.022*	"bind"	0.013*	"hormone"	0.010*	"nutrition"
0.022*	"transcription"	0.012*	"genetic manipulation"	0.009*	"triticum aestivum"
0.017*	"promoter"	0.012*	"growth hormone"	0.009*	"pollen"
0.015*	"evolution"	0.008*	"transgene"	0.009*	"tgf beta"
0.012*	"messenger rna"	0.006*	"luteinizing hormone"	0.008*	"cold tolerance"
0.010*	"peptide"	0.006*	"medicine "	0.008*	"cold acclimation"
T4 基因克隆技术		T5 阿尔茨海默症		T6 肿瘤	
0.040*	"molecular clon"	0.108*	"alzheimer disease"	0.039*	"cancer"
0.037*	"biosynthesis"	0.062*	"oxidative stress"	0.024*	"breast cancer"
0.037*	"purification"	0.054*	"amyloid precursor protein"	0.018*	"transport"
0.035*	"Escherichia coli"	0.016*	"neurodegeneration"	0.013*	"reveal"
0.033*	"gene clon"	0.014*	"hippocampus"	0.013*	"epigenetic inheritance"
0.025*	"clon"	0.014*	"neuron"	0.012*	"tumor"
0.025*	"accumulation"	0.012*	"pathology"	0.011*	"assay"
0.022*	"protein"	0.010*	"central nervous system"	0.011*	"gene family"
0.013*	"cdna"	0.010*	"tau"	0.009*	"beta catenin"
0.011*	"sequence"	0.010*	"superoxide dismutase"	0.007*	"growth"
T7 疫苗		T8 突触可塑性		T9 动脉粥硬化疾病	
0.023*	"antibody"	0.018*	"animal models"	0.015*	"activation"
0.019*	"vaccine"	0.018*	"synaptic plasticity"	0.014*	"phenotype"
0.016*	"transgene expression"	0.017*	"antioxidant"	0.014*	"calcium"
0.015*	"antigen"	0.016*	"blood pressure"	0.013*	"endothelial cell"
0.013*	"association"	0.015*	"behavior"	0.013*	"recognition"
0.011*	"growth factor"	0.014*	"nitric oxide"	0.011*	"muscle"
0.010*	"encode"	0.014*	"brain"	0.011*	"cerebrospinal fluid"
0.009*	"resistant"	0.011*	"knockout rat"	0.010*	"atherosclerosis"
0.009*	"immunization"	0.010*	"huntington disease"	0.009*	"fibroblast"
0.009*	"immunogenicity"	0.009*	"neural stem cell"	0.008*	"angiotensin ii"

图 3 基因工程领域的主题-词项分布（部分）

（2）主题强度和主题学科交叉度测度

通过计算主题强度和主题学科交叉度，本文对 2000 年至 2019 年各个主题的主题强度和主题学科交叉度进行比较分析，如表 2 所示。

表 2　基因工程领域各主题的强度及学科交叉度

序号	主题	主题强度	学科交叉度	序号	主题	主题强度	学科交叉度
T1	转录组测序技术	0.0816	0.3728	T12	转基因动物	0.0608	0.3944
T2	转基因药物	0.0266	0.4201	T13	基因疗法	0.0343	0.4298
T3	转基因作物	0.0331	0.4157	T14	生物遗传	0.0473	0.4244
T4	基因克隆技术	0.0725	0.3512	T15	植物抗病性	0.0654	0.2599
T5	阿尔茨海默症	0.0723	0.4578	T16	植物修复技术	0.0380	0.3642
T6	肿瘤	0.0383	0.4164	T17	细胞凋亡	0.0509	0.4355
T7	疫苗	0.0431	0.4095	T18	肌萎缩侧索硬化症	0.0402	0.4339
T8	突触可塑性	0.0384	0.4587	T19	生物多样性保护	0.0356	0.3857
T9	动脉粥硬化疾病	0.0411	0.4487	T20	生物体免疫反应	0.0423	0.4409
T10	额颞痴呆症	0.0356	0.4334	T21	非生物胁迫	0.0665	0.2089
T11	植物抗虫性	0.0362	0.4280				

通过计算得到基因工程领域主题强度阈值为 0.0476，获得 7 个热点主题，分别是“转录组测序技术”“基因克隆技术”“阿尔茨海默症”“转基因动物研究”“植物抗病性”“细胞凋亡”和“非生物胁迫”。通过计算得到基因工程领域主题学科交叉度阈值为 0.3995，获得 14 个学科交叉主题，分别是“转基因药物”“转基因作物”“阿尔茨海默症”“肿瘤”“疫苗”“突触可塑性”“动脉粥硬化疾病”“额颞痴呆症”“植物抗虫性”“基因疗法”“生物遗传”“细胞凋亡”、“肌萎缩侧索硬化症”和“生物体免疫反应”。

其中，主题强度与主题学科交叉度均超过阈值的主题有“细胞凋亡”与“阿尔茨海默症”，它们既是学科交叉主题也是热点主题，为热点交叉主题，如图 4 所示。

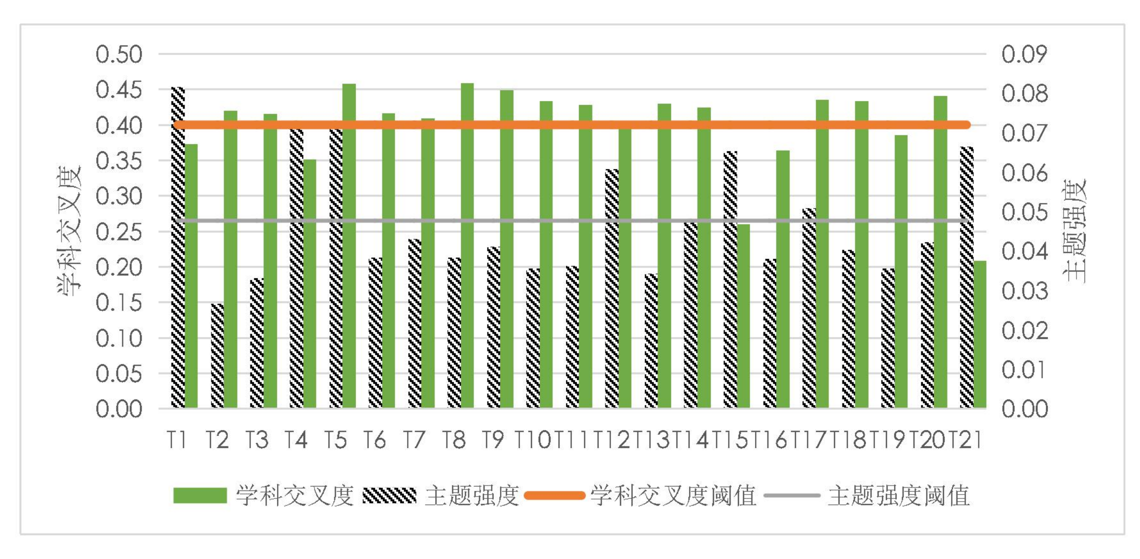
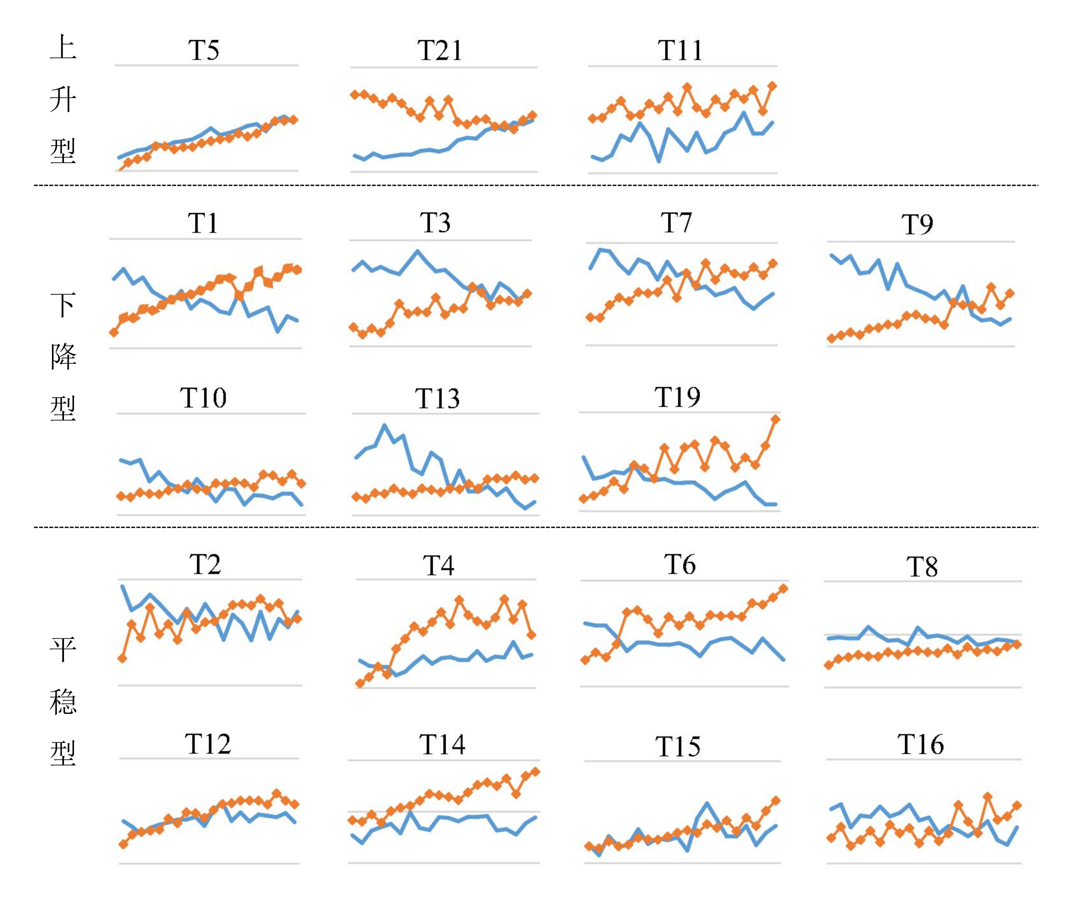


图 4 主题强度及学科交叉度分布图

3.3 主题演化与分析

以年为单位，获得文档所属年份，计算出在连续的时间窗口内各个主题的主题强度值及主题学科交叉度值，分别绘制基因工程主题强度和主题学科交叉度变化趋势图。通过观察各个主题在连续时间窗口下的主题强度及主题学科交叉度的变化情况，总结主题演化特征，将各个主题归结为上升型主题、下降型主题和平稳型主题。



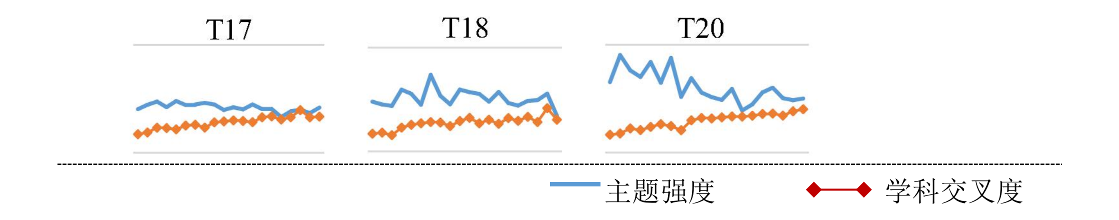


图 5　基因工程主题强度和主题学科交叉度变化趋势图

主题“T5 阿尔茨海默症”与“T17 细胞凋亡”，它们既是学科交叉主题也是热点主题。从图 5 可以看出，除“T21 非生物胁迫”外，其他主题学科交叉度基本呈现上升趋势，在基因工程研究领域，主题的学科交叉性正变得越来越强，跨学科合作愈加明显。3 个主题的强度在不断增加，7 个主题的强度呈现出下降趋势，其余 11 个主题的强度变化较小，呈平稳趋势。通过主题强度变化趋势将主题分为上升型主题、下降型主题和平稳型主题，综合主题学科交叉度，对重点主题进行分析。

（1）上升型主题

图 5 显示，上升型主题有 3 个，分别是 T5 阿尔茨海默症、T11 植物抗虫性和 T21 非生物胁迫。其中，“T21 非生物胁迫”为热点研究主题，“T5 阿尔茨海默症”为热点交叉主题。选择热点交叉主题“T5 阿尔茨海默症”重点分析如下：

20 年间“T5 阿尔茨海默症”主题强度一直很高，总体呈现上升趋势，说明该主题目前仍属于研究热点，受到学者们的高度关注。尽管主题强度分别在 2004 年、2009 年、2010 年下降，但下降幅度不大，并于次年及时回升，并未影响整体上升趋势。该主题的学科交叉度也呈现平稳上升趋势，表明越来越多的学科参与到阿尔茨海默症的研究当中，结合具体学科分析发现，Neurosciences 和 Biochemistry & Molecular Biology 对该主题研究贡献显著。目前，阿尔茨海默症的研究主要包括：发病机制、危险因素、诊断及治疗。这些研究问题的复杂性和综合性，需要打破学科壁垒进行多学科协同合作，可以合理预测未来随着老龄化人口数量逐渐上升，阿尔茨海默症主题强度仍然会不断上升，其研究热度也必将带动更多学科参与进来。

（2）下降型主题

2000 到 2019 年间，7 个下降型主题包括：T1 转录组测序技术、T3 转基因作物、T7 疫苗、T9 动脉粥硬化疾病、T10 额颞痴呆症、T13 基因疗法和 T19 生物多样性保护。其中，“T1 转录组测序技术”为热点主题，“T3 转基因作物”、“T7 疫苗”、“T9 动脉粥硬化疾病”、“T10 额颞痴呆症”和“T13 基因疗法”为学科交叉主题。选择学科交叉主题“T7 疫苗”重点分析如下：

在人类不断和疾病斗争的历史中，疫苗接种是有效消灭和控制传染性疾病的有效手段。随着科学技术的发展，针对各类疾病的疫苗被研发了出来，如乙肝疫苗、狂犬病疫苗等，其研发技术也在不断完善。疫苗的主题强度在 2001 年达到巅峰，随后呈波折下降趋势，但在 2017 年出现明显拐点，呈上升趋势，由此可见，疫苗研究热度虽然呈下降趋势，但在 2017

年之后热度有所回升。该主题的学科交叉度呈上升趋势，在 2005 年学科交叉度为 0.4095，成为学科交叉主题，Immunology 和 Biochemistry & Molecular Biology 两个学科的学者一直在重点关注疫苗研究。结合 2019 年底出现并造成疾病大流行的新型冠状病毒，世界各国加大投入资金、人员等各方面力量，通过学科交叉和跨界融合，使用变革性技术有效推动疫苗研发，大大缩短疫苗研制周期，成功研发出多类型的新冠疫苗，该主题强度及主题学科交叉度在 2019 年都出现增长。可以预测，随着人们“以预防为主”的健康意识的觉醒和公共卫生事件的发生，疫苗研究的热度未来将会不断升高，有望成为研究热点。

（3）平稳型主题

平稳型主题有 11 个，包括：T2 转基因药物、T4 基因克隆技术、T6 肿瘤、T8 突触可塑性、T12 转基因动物、T14 生物遗传、T15 植物抗病性、T16 植物修复技术、T17 细胞凋亡、T18 肌萎缩侧索硬化症、T20 生物体免疫反应。其中，“T4 转基因克隆技术”、“T15 植物抗病性”、“T12 转基因动物”为热点主题，“T2 转基因药物”、“T6 肿瘤”、“T8 突触可塑性”、“T11 植物抗虫性”、“T14 生物遗传”、“T18 肌萎缩侧索硬化症”、“T20 生物体免疫反应”为学科交叉主题，“T17 细胞凋亡”为热点学科交叉主题。选择热点学科交叉主题“T17 细胞凋亡”重点分析如下：

细胞凋亡指正常细胞在经过生理性或病理性的刺激之后由基因控制主动性死亡的过程。细胞凋亡主题强度一直处于较高的水平，且趋势平稳；主题学科交叉度稳中带升，意味着该主题一直是学者们关注的焦点，被应用于多种领域。由于细胞凋亡过程中整个细胞会生成含有细胞器、细胞核及细胞质的凋亡碎片的突起，然后被其他细胞吞噬，学者们根据这一特点，进一步了解生物体内的细胞，更是为医学、发育、畜牧业等领域带来了崭新的研究方向。由此可以看出，细胞凋亡研究在学科领域应用广泛，帮助多个学科攻克难题。但迄今为止，细胞凋亡的检测方法和凋亡途径等仍未被研究透彻，为了更彻底地了解细胞凋亡机制，更有效地治疗疾病，相关学科的学者们将持续对该主题保持关注，尝试对细胞凋亡进行更深入的研究。

4 结语

本研究对基因工程领域主题内容进行了挖掘提取，利用 LDA 主题模型生成的主题-词项概率分布确定出 21 个重要主题，随后基于主题强度和主题学科交叉度阈值识别出了基因工程热点交叉主题并进行主题分析。对文档主题按照时间进行划分，通过计算各个主题在不同且连续的时间窗口内的强度和学科交叉度，绘制基因工程研究主题强度和学科交叉度变化趋势图，进而获得主题演化趋势。本研究可以得出以下结论：

（1）识别学科领域热点交叉主题。使用 LDA 模型可以快速获取、识别基因工程领域 21 个重点主题，同时，结合 Rao-Stirling 指数测度该领域主题学科交叉度可从海量文献中快速发现基因工程领域 2 个热点交叉主题，该方法对其他领域的热点学科交叉主题研究同样具有适用性。

（2）主题强度具有动态变化性。从主题强度来看，基因工程领域热点研究主题包括：

“转录组测序技术”“基因克隆技术”“阿尔茨海默症”“转基因动物研究”“植物抗病性”“细胞凋亡”和“非生物胁迫”。受生物、信息技术不断发展或突发疾病及政治等因素影响，基因工程领域在不同时间窗口主题热度则会发生改变，相关学者将增加或减弱对主题的关注度。

（3）基因工程领域学科交叉融合程度进一步加深。从学科交叉度来看，大部分主题学科交叉度呈现上升趋势，当前基因工程领域愈发重视学科交叉研究。其中，“阿尔茨海默症”主题强度及学科交叉度持续上升，越来越多的学科投入研究，可进一步开展深度研究；“细胞凋亡”主题强度和学科交叉度平稳处于较高水平，从微观水平揭示生命的奥秘持续吸引着众多学者的关注。

本文仍存在一定局限性，在根据困惑度的计算结果得到最优主题数目时，存在由于主题数目过多导致主题辨识度偏低的风险；通过使用 LDA 主题模型获得基于每个主题下的概率前十的词项，在归纳总结后的命名结果并不能够完全概括主题下的所有内容，会存在一定的偏差；对基因工程领域主题的分析受限于时间和专业能力，未来可以通过阅读文献与采访领域内专家，进一步加深热点交叉主题的分析。

参考文献：

[1] 吴朝晖,赵婀娜.以学科交叉融合服务国家战略需求[N].人民日报,2020-11-04(012).

[2] 贾夏利.基于主题共现的交叉学科知识融合研究[D].北京:中国科学院大学(中国科学院文献情报中心),2022.

[3] 朱世琴,范丹丹,苟文静.学科交叉与知识扩散相关性研究——以基因工程领域为例[J].现代情报,2022,42(09):121-131.

[4] 李春景,刘仲林.现代科学发展学科交叉模式探析——一种学科交叉模式的分析框架[J].科学学研究,2004(03):244-248.

[5] 马费成.关注学科热点 透视学术进步[J].情报资料工作,2022,43(01):13-14+22.

[6] 许海云,尹春晓,郭婷等.学科交叉研究综述[J].图书情报工作,2015,59(05):119-127.

[7] Chi R, Young J. The interdisciplinary Structure of Research on intercultural Relations: A Co-citation Network Analysis Study[J].Scientometrics,2013,96(1): 147-171.

[8] Adams J, Light R. Mapping Interdisciplinary Fields: Efficiencies, Gaps and Redundancies in HIV/AIDS Research[J].Plos One,2014,9(12):e115092.

[9] Vugteveen P, Lenders R, Van den Besselaar P. The dynamics of interdisciplinary research fields: the case of river research[J]. Scientometrics, 2014, 100(1):73-96.

[10] Xu H, Guo T, Yue Z, et al. Interdisciplinary topics of information science: a study

based on the terms interdisciplinarity index series[J]. Scientometrics, 2016, 106(5):583-601.
[11]杜丽君.学科交叉视角下的信息检索研究主题演化分析——以情报学和计算机科学为例[J].信息技术与信息化,2020(01):178-183.
[12]隗玲,许海云,刘春江等.技术领域主题发现研究——以基因工程疫苗领域为例[J].数字图书馆论坛,2017(01):37-45.
[13]罗瑞,许海云,刘亚辉.基于结构熵的科学突破主题识别——以基因工程疫苗领域为例[J].情报理论与实践,2021,44(05):106-114+99.
[14]Blei D M. Probabilistic topic models[J]. Communications of the ACM, 2012, 55(4):77-84.
[15]Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of machine Learning research, 2003(03):993-1022.
[16]张斌.交叉学科主题探究:从主题聚类视角[J].情报科学,2020,38(10):49-55.
[17]陈琼,朱庆华,闵华等.基于领域主题的学科交叉特征识别方法研究——以医学信息学为例[J].现代情报,2022,42(04):11-24.
[18]韩正琪,刘小平,寇晶晶.基于 Rao-Stirling 指数和 LDA 模型的领域学科交叉主题识别——以纳米科技为例[J].情报科学,2020,38(02):116-124.
[19]Silva F N , Rodrigues F A , Oliveira O N , et al. Quantifying the interdisciplinarity of scientific journals and fields[J]. Journal of Informetrics, 2013, 7(2):469-477.
[20]Leydesdorff L, de Moya-Anegón, F, Guerrero-Bote, V P. Journal Maps, Interactive Overlays, and the Measurement of Interdisciplinarity on the Basis of Scopus Data (1996-2012)[J]. Journal of the Association for Information Science & Technology, 2013, 66(5):1001-1016.
[21]孟祥保.图书情报学交叉融合与发展——基于国外 35 种核心期刊的引文分析[J].图书情报知识,2012,(5):50-58.
[22] Agarwal R. On the intellectual structure and evolution of ISR[J]. Information systems research, 2016, 27(3):471-477.
[23] 杨瑞仙,姜小函.从学科和期刊的引证视角看交叉学科的知识结构和演化问题——以图书情报学科为例的实证研究[J].图书情报工作,2018,62(05):30-39.
[24]Carley, S, Porter, A L. A forward diversity index[J]. Scientometrics. 2012, 90(2), 407-427.
[25]Levitt J M , Thelwall M , Oppenheim C . Variations between subjects in the extent to which the social sciences have become more interdisciplinary[J]. Journal of the Association for Information Science &

Technology, 2011, 62(6):1118-1129.
[26] 曹嘉君,王曰芬,陈盛之等.多学科交叉综合的研究领域内学科间分布状态与演化研究[J].情报学报,2020,39(05):459-468.
[27]Deng S L, Xia S D. Mapping the interdisciplinarity in information behavior research: a quantitative study using diversity measure and co-occurrence analysis[J]. Scientometrics, 2020, 124(4):489-513.
[28] 吴查科,王树义.基于 LDA 的国内图书馆学研究主题发现及演化研究[J].新世纪图书馆,2019(07):90-96.
[29] 单斌,李芳.基于 LDA 话题演化研究方法综述[J].中文信息学报,2010,24(06):43-49,68.